eHealth Traffic Detection and Classification Using Machine Learning Techniques

Monika Grajzer
Telcordia Poland
Applied Research Center
Umultowska 85, 61-614 Poznań
Email: mgrajzer@telcordia.com

Piotr Szczechowiak
Telcordia Poland
Applied Research Center
Umultowska 85, 61-614 Poznań
Email: pszczech@telcordia.com

***Abstract*—With the growing number of available eHealth applications, the amount of eHealth traffic transmitted through communication networks increases significantly. This implies that network mechanisms must provide Quality of Service (QoS) assurances to support these new applications. In order to improve network performance, there is a need to develop new QoS methods that would properly detect and classify eHealth traffic. In this paper we present a selection of machine learning - based traffic classification methods in the context of eHealth services provisioning. We also present a mapping of eHealth application classes to appropriate QoS classes. Finally we propose an eHealth-aware approach, which can perform real-time traffic classification. In this technique the packet content is not inspected and at the same time the privacy of transmitted information is preserved.**

***Index Terms*—eHealth applications, traffic classification, flow analysis, machine learning;**

I. Introduction

High capacity and throughput of current telecommunication networks make it possible to provide remote eHealth services to users, no matter if they are at home or on the move. Telemedicine applications are gaining popularity and the amount of eHealth traffic transmitted through communication networks increases significantly. At the same time, network mechanisms must provide Quality of Service (QoS) assurances to a whole range of different eHealth applications. Appropriate service levels should be guaranteed for simple consulting services as well as life-critical clinical telemedicine applications.

The increasing demand to ensure appropriate bandwidth, maximum delay and jitter for telemedicine applications is very challenging for current network QoS mechanisms. eHealth services have diversified demands and each application type requires different treatment [1]. Existing QoS solutions [2], [3] were designed to support generic types of applications and have not been tuned to address e-Health needs. Current methods have difficulties in detecting eHealth traffic and cannot provide proper classification of medical applications. All these problems have a significant impact on the reliability of eHealth services. In order to improve network performance there is a need to develop new QoS methods that would properly detect and classify eHealth applications at the edge of the network.

We can identify several challenges in the design of QoS classification mechanisms for eHealth traffic. This kind of traffic is very often related to time-critical applications, where delays should be kept to minimum. Such behaviour requires real-time operation of QoS classification algorithms. Moreover, early detection of traffic type is necessary to allow proper handling by the network nodes. Since new eHealth applications are constantly emerging, the classifier should also have the ability to recognise previously unknown traffic. Finally the classification is usually performed on an encrypted traffic, which makes it difficult to asses the packet content. An ideal method for eHealth traffic classification should address all the above design challenges.

In this paper, we review different traffic classification techniques in the context of eHealth services provisioning. We propose the mapping of eHealth application classes to QoS classes and point out which e-Health traffic characteristics are the best metrics for the classification algorithms. We also propose an eHealth-aware classification approach, which is based on Machine Learning (ML) techniques. It takes advantage of IP traffic classification based on statistical properties of a flow. It applies complex classification techniques, where decision is made through the multi-criteria reasoning without looking deep into the packet content. By employing this approach the privacy of the data is intact and classification can be performed at the edge of the network in real time to guarantee proper handling of eHealth traffic by each network node.

II. Related work

Traffic classification is an important aspect of every QoS solution and is usually performed at the edge of the network, where the application traffic originates. In the literature on QoS solutions for eHealth traffic, we can find QoS mechanisms that investigate priority assignment and scheduling techniques for eHealth applications [4]. However, these techniques cannot distinguish medical applications and take an assumption that eHealth traffic can be easily detected. Such an approach may lead to incorrect classification of eHealth data and decreased quality of service. Therefore ehealth tarffic classification is a subject to our research.

In general, there are several methods, which address the traffic classification problem in the communication networks

 ISBN: 978-1-61208-179-3

[2], [3]. The basic and most common methods are focused on the evaluation of the QoS related fields in the IP packet header. They look at the so called "5 tuple" - a set of packet's source IP address, destination IP address, source and destination TCP/UDP port number and layer 4 protocol type [2]. Although such classification is simple, fast and differentiates basic networking services (e.g., voice service from data service), it very often mis-classifies the traffic or puts traffic from diversified applications into one category. Therefore it was argued that such a simple method is not enough to perform complex classification tasks [2], [3], [5].

Additional information about a network flow can be obtained by analyzing the information contained in the packet payload. This approach, denoted as Deep Packet Inspection (DPI) [2], investigates application-specific packet metrics, which can significantly increase classification success rate. The main drawbacks of this approach are problems with accessing encrypted messages and high computational complexity of payload analyzing algorithms (hardware implementations needed). Moreover DPI techniques require previous knowledge about application-specific parameters, which have to be updated for every new application. Since medical data is usually encrypted and requires real-time packet processing the applicability of DPI methods to classify this type of traffic is rather limited.

The techniques described above (port numbers, DPI) have limited practical relevance in case of eHealth applications, but they can still serve as a reference (ground-truth) for more advanced methods. New solutions might be based on machine learning classifiers, which are capable of making decisions based on the observation of the traffic flow features like packet lengths and packet inter-arrival times. This makes them particularly suitable for the classification of e-Health traffic.

Previous works on ML traffic classification [3], consider both supervised and unsupervised learning approaches. Both techniques usually require a training phase to precede the classification phase. Supervised learning techniques are particularly suitable to solve classification problems, whereas unsupervised learning techniques enable clustering of traffic flows into groups sharing common features. As such, they must be accompanied by the labelling algorithm that would assign particular applications to the identified clusters, which is challenging. The additional benefit of these methods is the ability to classify applications which are unknown. The examples of ML classifiers that can be used for QoS mechanisms are: J48 Decision Tree, K-Nearest Neighbourhood, Random Tree, Naive Bayes and the Neural Networks method [3], [5]. Although ML-based techniques have several features making them suitable to e-Health traffic classification, not all of them are applicable to this particular problem. It has been also presented that the accuracy of a single classifier is not good enough to classify different types of applications when early classification is required [5].

A selection of ML techniques and classifiers suitable to e-Health traffic is presented in Section IV. Based on this methods, we propose a solution, which combines different classifiers in order to maximize the performance and accuracy of eHealth applications classification.

TABLE I
EHEALTH APPLICATIONS CLASSIFICATION BASED ON THE REQUIRED QOS PARAMETERS

Service type	Bandwidth	Delay	Packet loss	Class
Teleconsultation	High	Low	Low	1
HPC services	High	Low	Min	1
HD video	Max	Low	Low	1
HD images	High	Medium	Min	2
Sensor data	Medium	Low	Min	3
Patient data	Medium	Medium	Low	4
e-learning	High	Medium	Low	4
Voice	Medium	Low	Medium	5

III. EHEALTH TRAFFIC CHARACTERISTIC

There are various types of medical applications and each of them has different requirements when it comes to quality of service parameters. The QoS measures have been well defined in the domain of communication services; however, they have not been well defined in the eHealth domain. eHealth service category introduces a more diversified set of traffic patterns with bandwidth, delay and reliability requirements very different from typical network flows. They also introduce some of the highest QoS requirements among all services provided over IP networks. Real-time teleconsulation services require high throughput and are susceptible to packet delays (speech and HD video transmission). Video streaming from endoscopes and other medical devices has similar requirements. Medical images transmission needs high bandwidth and a very low packet loss ratio. This is especially important in case of high resolution images (X-ray, MRI, USG), because distorted images may lead to a wrong diagnosis. Clinical telemedicine applications, medical simulation tools and image reconstruction programs require guaranteed data delivery and minimal packet loss. Data transmission from medical sensors is not delay tolerant and even minimal packet loss is unacceptable (e.g., heart rate monitoring sensors).

Before an appropriate traffic classification technique can be designed, there is a need to characterize different types of eHealth traffic and identify all its distinctive features. Based on this information applications can be grouped together into classes, which require similar service parameters and forwarding priorities. This is necessary to properly mark classified packets and introduce further QoS mechanisms in the network (e.g., scheduling).

In Table I, we present the result of our analysis of eHealth application types in the context of the required QoS parameters. The above classification distinguishes five basic classes of eHealth applications. The first class contains highest priority services, which have strict requirements regarding bandwidth and maximum packet delay (e.g. High Perfomance Computing - HPC services). The second class contains data transmission services, which need guranteed packet loss rates. The third

 ISBN: 978-1-61208-179-3

group of applications gathers services which have low packet loss and delay. The next class aims at high bandwidth provisioning and has some tolerance against packet delay. The last class targets VoIP connections in eHealth networks, which require low packet delay and jitter.

IV. Classification algorithms and traffic classifiers

Having in mind the requirements towards classification of e-health traffic, we have identified several ML-based techniques as particularly suitable to the analysis of e-Health traffic. As a general rule ML classifiers investigate multiple flow descriptors - called *features* - simultaneously and provide learning capabilities, which introduce adaptive behaviours. They need to be trained on the examples of traffic flows from different applications, and the proper learning process is crucial to the final performance of the classifier. Moreover, the performance is very often dependent on the type of the data and on the set of features selected for observation, which describe characteristics of given application. Therefore, not all well performing classifiers would be valuable during the analysis of e-Health traffic.

In order to identify ML classifiers, which are the most suitable for our research subject, we have defined a set of criteria that are driven by e-Health traffic characteristics:

- Real time operation – most techniques require the observation of a full traffic flow to provide good classification results. This approach imposes significant delay, which is not tolerable in our case. Therefore in our solution we need methods where the observation of a whole flow is not required.
- Low number of necessary packets – a decision regarding classification needs to be made on the shortest possible part of the flow.
- Ability to perform classification based on a randomly selected part of the flow – in many cases the beginning of a flow cannot be observed, and the classifier should still make an accurate decision.
- Small processing overhead – a lightweight solution is required, but higher overhead is acceptable in the learning phase, which is performed offline.
- Ability to classify currently unknown application types.
- Small number of required features.
- Ability to work with encrypted traffic.

Below, we present an overview of the ML methods which were selected based on the evaluation of the above criteria:

1) Simple K-Means method: This method, proposed in [6], is the unsupervised learning approach based on the Simple-K-Means algorithm. The main advantage of this method is that it only needs the first few packets of the traffic flow, which depict application's negotiation phase [6]. It is thus assumed that unique negotiation phase is the differentiator between applications. This method has also very small set of features limited to the investigation of packet lengths. As an unsupervised learning method, Simple-K-Means divides observed traffic flows into clusters. During the learning phase each cluster is associated with a set of related applications. The particular flow in the cluster is classified as belonging to the most prevalent application from this set. The most challenging aspect of this approach is to properly assign different applications to clusters, so that given application is dominating in at least one of them and thus can be selected as a result of the classification. Although real-time classification is possible with this approach, difficulties might occur when the beginning of the flow is lost.

2) Multiple Sub-Flows method: The authors in [7] propose a supervised learning solution, which allows classification a flows based on the observation of N consecutive packets from any part of the flow. This feature is an important asset of the method. During the training process sub-flows of length N are extracted from the original flow, which represent parts with diversified characteristics. The classifier is trained on the sub-flows instead of the original flow. Therefore, the number of packets required for actual classification is relatively small (around 25), likewise is the number of necessary features. Capturing the start of the flow is not required. This method fulfils many of the identified criteria. Its disadvantage is however the inability to identify new application types.

3) Statictical protocol fingerprint method: This approach, presented in [8], analyses the flow and extracts its statistical properties, called protocol fingerprints, that would correspond to the behaviour of given protocols. This is performed in a training phase. Supervised learning - based classification is then performed by comparing those fingerprints to the statistical behaviour of the observed flow. On this basis particular protocols are identified. The method requires evaluation of only 3 features and enables real-time processing by observing first few packets of the flow. This approach is a promising technique with good performance. However when applied to e-Health traffic it may not always be possible to identify the application types correctly. This is because many applications would use several protocols in parallel, e.g. to transmit voice and data information separately.

4) Semisupervised classification method: Erman et al. [9] proposed a hybrid approach that takes advantage from both supervised and unsupervised ML techniques. The aim was to minimize the problem with proper labeling of clusters, which are the result of unsupervised techniques. Although this method requires the observation of full flows, it can be still valid in our case since it provides a unique ability to classify unknown application types. In this approach, both labeled and unlabeled flows are used in the training phase. Unsupervised method is exploited to form clusters whereas labeled flows in the cluster provide a way to map the cluster to the particular application type. The classifier, i.e., supervised technique, is then used to map unlabeled flows to one of the clusters/applications. An advantage of this method is reduced processing overhead in the learning phase.

5) Multi-classification: The authors in [5] observed that a significant number of network traffic classifiers performs well when applied to full flows. In order to work with parts of the flows, more sophisticated classification methods

 ISBN: 978-1-61208-179-3

TABLE II
SELECTION OF TRAFFIC CLASSIFICATION METHODS VS. CRITERIA IDENTIFIED FOR E-HEALTH TRAFFIC

	Real time operation	Number of packets	Flow beginning can be skipped	Processing overhead	Unknown applications	Small number of features	Encrypted traffic
5-tuple (Ports)	Yes	Low	Yes	Low	No	N/A	No
DPI	No	Low	Yes	Very High	No	N/A	No
Simple K-Means	Yes	Low	No	Low	No	Yes	Yes
Multiple Sub-Flows	Yes	Low	Yes	Average	No	Yes	Yes
Protocol fingerprint	Yes	Low	No	Average	No	Yes	Yes
Semisupervised	No	High	No	Average	Yes	Yes	Yes
Multi-classification	Yes	Very low	No	Moderate-High	No	Not clear	Yes

are required. Based on these observations they proposed a multi-classifier approach. In this method several classification techniques are combined to work in parallel. Classification is made based on the observation of a very short fragment of the flow ($<$10 packets). Although each classifier alone would perform rather poorly under such conditions, the appropriate combination of outputs from standalone classifiers can significantly increase the performance. However, this fast traffic classification method is performed at a cost of increased processing overhead.

V. OUR APPROACH

Table II presents the comparison of different QoS classification techniques based on the criteria specified for e-Health traffic. It can be observed that none of the standalone classifiers would be able to fulfil all the requirements that we have identified for eHealth traffic. Moreover, the performance of these methods, when applied to classification of eHealth traffic, should be verified through experimental results on real-life traffic streams. Therefore, in our future research we will focus on implementing different multi-classification methods that would take advantage of the standalone classifiers described in the previous chapter. We will investigate different combination techniques to achieve the optimal set of features and the best classification accuracy for any type of eHealth traffic.

The eHealth application classes proposed in Section III will be used by the ML-based classification techniques as the base for assigning clusters into appropriate flow groups. Such an approach will allow straightforward mapping of eHealth application classes into the appropriate QoS classes. In this way, the results of eHealth traffic classification could be directly used by the packet scheduling mechanisms used in existing QoS architectures (e.g., Diffserv, Interserv).

VI. CONCLUSIONS

The main difficulties in eHealth services provisioning are connected with the reliability and privacy issues of personal data transmissions over public networks. Ubiquitous eHealth service category poses the most stringent performance requirements to Internet technology and network systems in terms of quality of service due to its nature of life and liability. Current methods for QoS provisioning over IP networks were not designed to guarantee reliable transfer of data for eHealth applications. The main problems lay in proper eHealth traffic detection and classification in order to assign packets to appropriate QoS classes.

This paper presented an overview of traffic classification methods, which might be applicable to different eHealth applications. It proposed a basic mapping of eHealth application classes to appropriate QoS classes and also proposed machine learning - based traffic classification techniques for real-time packet flows. The proposed methods are able to address most of the challenges of eHealth traffic classification and does not require any packet payload inspections. In this way the privacy of the transmitted information can be preserved.

ACKNOWLEDGMENT

This work is partially supported by the Polish Ministry of Science and Higher Education under the grant agreement no. 543/N-COST/2010/0 ”Traffic analysis in eHealth networks”.

REFERENCES

[1] L. Skorin-Kapov and M. Matijasevic, “Analysis of qos requirements for e-health services and mapping to evolved packet system qos classes,” *International Journal of Telemedicine and Applications*, vol. 2010.

[2] J. Evans and C. Filsfils, *Deploying IP and MPLS QoS for multiservice networks: theory and practice*, ser. The Morgan Kaufmann Series in Networking, R. Adams, Ed. Morgan Kaufmann Publishers is an imprint of Elsevier, 2007.

[3] T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” *Communications Surveys & Tutorials, IEEE*, vol. 10, no. 4, pp. 56–76, 2008.

[4] P. Koutsakis, “Guaranteed bandwidth allocation and qos support for mobile telemedicine traffic,” *2008 IEEE Sarnoff Symposium*, pp. 1–5, 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4520075

[5] A. Dainotti, A. Pescap, and C. Sansone, “Early classification of network traffic through multi-classification,” in *Traffic Monitoring and Analysis*, ser. Lecture Notes in Computer Science, J. Domingo-Pascual, Y. Shavitt, and S. Uhlig, Eds. Springer Berlin / Heidelberg, 2011, vol. 6613.

[6] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, “Traffic classification on the fly,” *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 2, pp. 23–26, 2006.

[7] T. Nguyen and G. Armitage, “Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks,” in *Proceedings. 2006 31st IEEE Conference on Local Computer Networks*. IEEE, 2006, pp. 369–376.

[8] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, “Traffic classification through simple statistical fingerprinting,” *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 1, pp. 5–16, 2007.

[9] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, “Semi-supervised network traffic classification,” in *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. ACM, 2007, pp. 369–370.